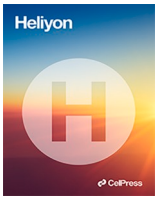
Heliyon 10 (2024) e31873

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A decentralized federated learning-based cancer survival prediction method with privacy protection

Hua Chai [a], Yiqian Huang [a], Lekai Xu [a], Xinpeng Song [a], Minfan He [a], Qingyong Wang [b,c,*]

[a] School of Mathematics and Big Data, Foshan University, Foshan, 528000, China
[b] School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, 230036, China
[c] Anhui Provincial Engineering Research Center for Agricultural Information Perception and Intelligent Computing, Hefei, 230036, China

ARTICLE INFO

Keywords:
Privacy protection
Federated learning
Survival prediction
Gene selection

ABSTRACT

Background: Survival prediction is one of the crucial goals in precision medicine, as accurate survival assessment can aid physicians in selecting appropriate treatment for individual patients. To achieve this aim, extensive data must be utilized to train the prediction model and prevent overfitting. However, the collection of patient data for disease prediction is challenging due to potential variations in data sources across institutions and concerns regarding privacy and ownership issues in data sharing. To facilitate the integration of cancer data from different institutions without violating privacy laws, we developed a federated learning-based data integration framework called AdFed, which can be used to evaluate patients' survival while considering the privacy protection problem by utilizing the decentralized federated learning technology and regularization method.

Results: AdFed was tested on different cancer datasets that contain the patients' information from different institutions. The experimental results show that AdFed using distributed data can achieve better performance in cancer survival prediction (AUC = 0.605) than the compared federated-learning-based methods (average AUC = 0.554). Additionally, to assess the biological interpretability of our method, in the case study we list 10 identified genes related to liver cancer selected by AdFed, among which 5 genes have been proved by literature review.

Conclusions: The results indicate that AdFed outperforms better than other federated-learning-based methods, and the interpretable algorithm can select biologically significant genes and pathways while ensuring the confidentiality and integrity of data.

1. Introduction

Refining the computational approach to accurate evaluate patients' survival is one of the pivotal objectives in cancer precision medicine [1]. Precise prediction models can aid clinicians in selecting optimal follow-up interventions based on individual patient characteristics. In previous studies, Zhou employed the support vector machine (SVM) classifier to predict ovarian cancer recurrence [2], while Pieretti utilized the logistic regression model to investigate potential biomarkers influencing cancer patient heterogeneity [3]. With the advancement of deep learning technology, various neural network-based approaches have been proposed for cancer

* Corresponding author. School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, 230036, China.
E-mail address: wang@cnu.ac.cn (Q. Wang).

https://doi.org/10.1016/j.heliyon.2024.e31873
Received 18 July 2023; Received in revised form 18 May 2024; Accepted 23 May 2024
Available online 23 May 2024

research. For instance, Wang utilized the convolutional neural network to extract CT image features from patients [4], while Hua et al. employed a denoising autoencoder to reconstruct the low-dimensional features of the high-dimensional gene expression data for evaluating cancer prognosis [5].

These computational techniques necessitate extensive data sets to train a robust model that avoids overfitting. However, cancer patient data is often fragmented due to collection by various institutions worldwide. To effectively utilize this distributed information for cancer research, some challenges related to information sharing cooperation must be considered. Personal medical information is highly sensitive and protected by the General Data Protection Regulation (GDPR), which requires consent from owners before sharing [6]. Hence, improving the utilization of medical data from multiple parties while ensuring data security and cooperation fairness remains a challenge for institutions seeking to collaborate. The researchers must rigorously prevent data leakage when utilizing medical data, and strike a balance between acquiring more information and safeguarding patients' privacy.

The utilization of distributed medical data for cancer research poses several challenges. Firstly, the inclusion of sensitive and personally identifiable information in medical data. Establishing robust measures such as data sharing agreements, anonymization techniques, and secure data transfer protocols is imperative to safeguard patient privacy and prevent any potential breaches. Secondly, accessing distributed medical data entails compliance with diverse legal, ethical, and regulatory considerations. Different institutions may have distinct policies and procedures governing data sharing and access. In this study, we focus on addressing the primary problem by proposing a machine learning technology-based framework for integrating and training data, enabling cancer research to leverage larger patient datasets while considering the privacy protection.

Traditional methods of protecting medical data privacy mainly rely on anonymous technologies, such as the (a, k)-anonymity algorithm developed by Li et al. [7], which aims to safeguard patients' privacy by concealing the link between their personal information and medical records. However, these techniques that obscure individual attributes in data are vulnerable to chaining attacks. Compared to anonymized privacy models, differential privacy computation is more resistant to privacy attacks and provides provable guarantees. Li proposed a private partitioning algorithm for differential privacy that reduces computational overhead and query errors, which was applied in electronic medical record queries [8]. Raisaro employed a combination of differential privacy protection and bidirectional decryption methods to safeguard genomic data against any potential attackers [9]. Nevertheless, methods based on differential privacy protection still necessitate users to upload their data onto a shared server. In some cases, participating parties may be unwilling to share their encrypted data with others.

Federated learning (FDL) is another widely-used solution for utilizing datasets collected from different institutions without centralizing or sharing training data [10], thereby protecting user privacy through the exchange of encrypted processed parameters while preventing access to source data by unauthorized parties. Various FDL-based machine learning methods have been developed in healthcare domains. Lee developed a federated privacy prediction platform that enables cross-institutional matching of similar patients in different hospitals for the purpose of sharing treatment information [11]. Chen established a FedHealth framework utilizing federated transfer learning technology to aggregate patient data for disease research while maintaining privacy and security [12]. Elayan proposed a distributed deep federated learning algorithm to collaboratively integrate health data collected from various organizations [13]. FedProx (Federated Proximal) is a commonly used federated learning approach that incorporates a proximal term into the loss function, promoting local models' similarity with the global model and preventing overfitting while enhancing generalization performance [14]. FedAvg (Federated Averaging) is a distributed machine learning approach that ensures privacy protection by aggregating model weights from multiple devices after local training, and subsequently updating the global model with these aggregated weights [15]. Additionally, federated learning has found extensive applications in various domains such as wind power prediction [16] and mitigation of false data injection attacks [17].

However, mitigating bias in a federated-learning-based cancer research framework poses a formidable challenge due to the heterogeneity of multi-source data. The optimization of hyperparameters, such as learning rate and regularization strength, can profoundly impact model performance and is crucial for enhancing both accuracy and efficiency. Furthermore, addressing the interpretability of features in privacy prediction models can help researchers identify the potential cancer-related genes.

To tackle these challenges, we have developed a crypto machine learning framework called Adaptive Decentralized Federated Learning (AdFed), which leverages decentralized federated learning technology and regularization methods to construct accurate and robust models for cancer survival prediction while safeguarding the privacy of data owners.

The proposed AdFed can provide interpretable results for identifying key factors that affect the survival of cancer patients. In this study, AdFed was tested on four different cancer datasets, each containing patient information collected from more than three institutions. The experimental results demonstrate that AdFed achieves a 2.7 % increase in AUC compared to the advanced encryption method. Furthermore, biological analysis indicates that AdFed can generate biologically meaningful outcomes for identifying cancer-related biomarkers. The remainder of this paper is organized as follows: Section 2 describes the related works, including the cancer datasets used, related work, and methodological details. In Section 3, we present the results comprising method evaluation and biological analysis. Finally, in Section 4, we provide conclusions and discuss future work.

2. Materials and methods

2.1. Dataset

In this study, four types of cancer patients were collected from TCGA (https://tcga-data.nci.nih.gov/tcga/) and GEO (https://www.ncbi.nlm.nih.gov) databases for method evaluation (Table 1). Gene expression features in the cancer datasets underwent log transformation normalization, while missing values were imputed based on median values. The batch effect of all datasets was corrected

using the R package "*limma*" [18]. Patients with a survival time exceeding 3 years were classified as low-risk groups, while the remaining patients were divided into high-risk subgroups.

2.2. AdFed framework for data integration

The General Data Protection Regulation (GDPR) is a comprehensive data protection regulation implemented by the European Union (EU) to safeguard the privacy and personal data of individuals within the EU [19]. Researchers must obtain explicit and informed consent from individuals for processing their personal data, specifically medical data, as required by the GDPR. The purpose of data processing should be clearly defined and limited to stated research objectives. Researchers must ensure compliance with GDPR requirements when processing medical data for research purposes, particularly in cancer research settings. This ensures the protection of individuals' privacy rights while facilitating valuable advancements in scientific knowledge.

FDL is one of the widely-used solution for utilizing datasets collected from different institutions while considering the privacy protection. While most of the existing FD-based methods addresses privacy concerns by not sharing raw data, the central server still aggregates and updates the global model. This centralization may raise concerns regarding data control and potential bias in model updates if not properly managed. In this study, AdFed leverages decentralized federated learning technology and regularization methods to construct accurate and robust models for cancer survival prediction while safeguarding the privacy of data owners. AdFed incorporates robust mechanisms to ensure the security of data and fairness in cooperation when utilizing medical data from multiple parties. During the AdFed learning phase, each participating party is as sever to aggregation parameters one by one, which remains securely stored on local devices or servers, thereby minimizing the risk of unauthorized access or data breaches. This decentralized approach guarantees that sensitive medical data remains within the boundaries of respective organizations, thus enhancing overall data security.

Moreover, prior to participating in AdFed, the data from each party is typically subjected to anonymization or pseudonymization techniques in order to safeguard patient privacy. Anonymization eliminates personally identifiable information, while pseudonymization replaces identifying details with pseudonyms. By employing these methodologies, the confidentiality of patients' medical data is upheld, thereby mitigating the risk of re-identification. AdFed utilizes secure communication protocols for data transmission purposes as well; all exchanges between the central server and participating entities are encrypted to ensure utmost confidentiality and protection against unauthorized interception or tampering.

Furthermore, the distributed server or aggregator in AdFed plays a pivotal role in ensuring robust data security and equitable cooperation [20]. It should function as a trusted entity that adheres to stringent governance protocols and enforces data usage agreements effectively. The aggregator must have restricted access to sensitive data, while its operations should be transparent, auditable, and accountable. AdFed facilitates collaborative utilization of medical data from multiple entities while upholding stringent measures for data security, safeguarding patient privacy, and promoting fair cooperation. These mechanisms bolster trust and encourage organizations' active participation in initiatives aimed at sharing data, thereby fostering invaluable insights and advancements in medical research.

As shown in Fig. 1, AdFed comprises an aggregation and a local model optimization step, both of which are executed by the clients on each learning iteration. Unlike server-based FDL methods, AdFed is serverless as the clients can collaborate with one another without requiring coordination from a central server. Unlike the conventional approach of transmitting model updates to a central server, AdFed's sites can directly disseminate learning model parameters to authenticated members of the FDL learner pool through peer-to-peer communication links.

Our proposed approach can serve as a pivotal concept in the context of cancer survival prediction with privacy preservation. By leveraging decentralized federated learning, it becomes feasible to train predictive models while upholding the confidentiality of sensitive medical data. Cancer survival prediction frequently requires utilizing highly sensitive and confidential individual-level medical data. In order to address privacy concerns adequately, FedAvg ensures that raw data is not disclosed; rather, it aggregates and updates the global model on a central server. Nevertheless, if not effectively managed, this centralized approach may give rise to apprehensions about both data control and potential biases in model updates. Henceforth, our proposed method facilitates training predictive models directly on local devices or servers where the relevant data is stored without necessitating any sharing of raw information. By adopting this decentralized approach, we can ensure that sensitive patient information remains securely controlled by their respective healthcare institutions or organizations while preserving patient privacy.

Additionally, the data obtained from various healthcare institutions or research centers may exhibit heterogeneity in terms of patient demographics, treatment protocols, and genetic variations. To address this heterogeneity, our proposed method allows for the

Table 1
The statistical information about the four cancer data used in this study.

Colon		Head and Neck		Liver		Ovarian	
TCGA	249	TCGA	403	TCGA	351	TCGA	296
gse17538	232	gse41613	97	gse10141	80	gse14764	79
gse38832	122	gse42743	57	gse14520	221	gse17260	108
				gse54236	81	gse18520	53
						gse32062	260
						gse63885	75
Total	603	Total	557	Total	733	Total	871

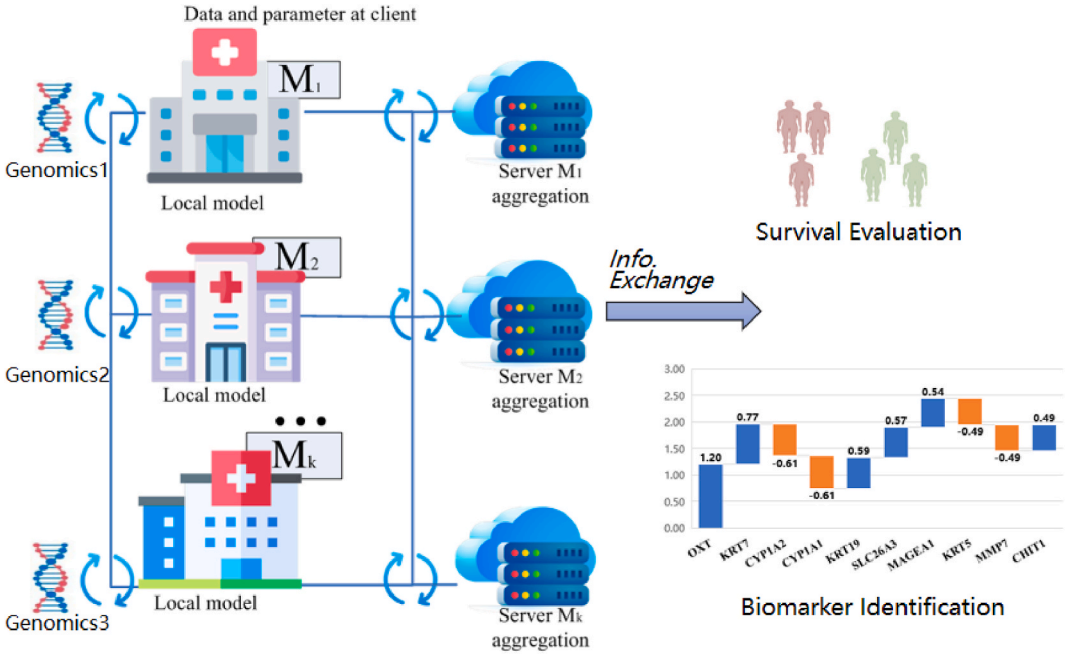


Fig. 1. The architecture of AdFed proposed in this study. AdFed enables each node in the flowchart to perform local computations independently and communicate with other nodes for information exchange and model updates. Each node takes turns serving as a parameter aggregation server, which aggregates the global model fed back to client devices for validation and inference.

training of local models using institution-specific data to capture the unique characteristics of each dataset.

2.3. Method details

In AdFed, the clients employ an ad-hoc aggregation step to incorporate information collected from local neighborhoods into the local model adaptation process, typically utilizing consensus methodologies. The aim of AdFed is to train and distribute models across multiple devices while ensuring the privacy of users. The loss function for federated learning can be expressed as follows:

$$min_w \sum_{k=1}^{K} (f_k(w;x) - y_k) \tag{1}$$

where w represents the model parameters, $f_k(w)$ denotes the number of samples on the k-th device, and n is the total number of samples across all devices.

Assuming an equation that minimizes $L\theta(w)$ demonstrates the optimization in AdFed for federated learning, with the objective of identifying values of w that minimize $L(w)$.

AdFed iteratively updates the values of w by following the gradient of the loss function with respect to θ, moving in a direction that leads to a decrease in loss until reaching its minimum. Here is an optimized example of applying the gradient descent algorithm to equation (1):

$$\theta(w)_{t+1} = \theta(w)_t - \alpha \nabla L(\theta(w)_t) \tag{2}$$

In equation (2), w_{t+1} denotes the updated value of w after one iteration of the algorithm, while w_t represents its current value. α stands for the learning rate or step size, and ∇L(w_t) refers to the gradient of the loss function with respect to w evaluated at its current value. equation (3) operates by iteratively adjusting the hyperparameter value of θ in the direction opposite to that of the gradient, scaled by a learning rate α, until convergence is achieved.

$$w^{k*} = argmin_{w^k} \frac{n_k}{n} \sum_{k=1}^{K} f_k(\theta(w^k)) \tag{3}$$

Hence, AdFed aims to solve equation (4) and obtain an optimized w:

$$w = \frac{1}{K} \sum_{k=1}^{K} (w^k) \tag{4}$$

where model has K clients, the parameter t controls the stability of the update model parameter w, t contains the neighbors of client k at round t. It should be noted that, each client in AdFed can either defer model aggregation until their neighbors complete local model optimization (synchronous implementation), or apply model aggregation as soon as they complete their own model optimization regardless of neighbor status (asynchronous implementation).

The model's ability to select features is crucial for understanding the significant gene features that are associated with patient survival. In AdFed, feature selection refers to the process of selecting a subset of features that provide the most informative value from multiple clients' available features. The chosen characteristics are subsequently consolidated at the central server to establish a universal model that can effectively generalize on unfamiliar data. Additionally, feature selection has the potential to diminish AdFed's communication expenses and enhance model performance by eliminating superfluous or irrelevant features. The selection feature set ($S(q,k)$ for the qth feature of the k-*th* party is given in equation (5):

$$S(q,k) = [[w_2^{k(q)}]]_1 \tag{5}$$

AdFed employed the integration of L1 and L2 norm loss functions in federated learning to select significant gene features. Specifically, each client utilized L2 regularization while their parameters were aggregated and the server selected key features with L1 regularization. The adoption of L2 regularization by clients facilitated the smooth aggregation of results, while the server's utilization of L1 regularization generated sparse outcomes through zero-valued coefficients. In this way, our feature selection approach effectively handles noisy and outlier features. Here we list the symbol introduction in Table 2.

2.4. AdFed algorithm optimization

Algorithm 1. demonstrates that the optimization of the AdFed algorithm is conducted on decentralized devices, where the parameters of each device are optimized to participate in every round of training. Each device trains the model based on its local data and computes an update. Updated information is collected and combined to form a global model, which is then refined through iterative processes until convergence. To ensure the privacy and security of data during this process, optimization algorithms may employ additional techniques such as differential privacy and encryption. The algorithm of AdFed is given as follows:

Table 2
The summary of the symbols used in our research.

Symbol	Explanation	Symbol	Explanation
w	Model parameter	b	Batch number
θ	Hyperparameter	α	Learning steps
L	Loss function	K	Clients number
n	Sample number		

Algorithm 1
AdFed

Procedure AdFed():

Initialize parameters

for k **in** K **do**
  for each round in t=1,2,…,T do
    w_{t+1}^k = ClientUpdate(k,);
    receive{[[w_t^k];
    The coordinator aggregates k the received model weight w:
    $[[\overline{w}_{t+1}^k]] \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} [[w_{t+1}^k]]$
  Send Enc([[w^k]])
  end for
end for
Proceduce ClientUpdate(t,w^k):
Dec [[w^k]]
ϑ ←Mini-batches of size B
for (each local epoch j=1,2,…,E) **do**
  for batch **do**
    Equation (2)
    Update $w_{b+1}^k = w_{b,j}^k - \alpha(\nabla(w_b^k))$
  end for
end for
Obtains the local weight $w_{t+1}^k = w_{B,E}^k$
Send Enc ([[wk]])

When tuning the parameters for a federated learning method, several considerations come into play. Hyperparameters like learning rate, momentum, and batch size influence the convergence and performance of the model. These parameters are adjusted to ensure optimal model performance by 5-fold cross-validation. Additionally, we give a parameter list in the following: the number of training epochs for each party was chosen from [1000,5000,10000], batch size [32,64,128], and the learning rate was set [1E−1, 1E-2, …,1E−5], and the momentum was set from [0.1,0.5,0.9].

Table 3
The predicted performance obtained by different methods in four cancers.

Data	Metric	RF	XGBoost	SVM	FedProx	FedAvg	AdFed
COAD	ACC	0.460	0.526	0.460	0.603	0.708	0.715
	AUC	0.519	0.520	0.537	0.601	0.627	0.635
	F1	0.381	0.223	0.471	0.523	0.732	0.744
HNSC	ACC	0.582	0.542	0.475	0.402	0.527	0.556
	AUC	0.520	0.480	0.511	0.396	0.537	0.542
	F1	0.423	0.204	0.300	0.293	0.272	0.293
LIHC	ACC	0.464	0.466	0.456	0.536	0.529	0.542
	AUC	0.446	0.424	0.520	0.541	0.565	0.576
	F1	0.267	0.220	0.304	0.371	0.389	0.409
OV	ACC	0.433	0.398	0.391	0.532	0.593	0.608
	AUC	0.522	0.510	0.523	0.535	0.584	0.601
	F1	0.520	0.511	0.525	0.540	0.601	0.648
Ave	ACC	0.485	0.483	0.447	0.518	0.589	0.605
	AUC	0.502	0.483	0.523	0.518	0.578	0.588
	F1	0.397	0.290	0.400	0.432	0.498	0.524

3. Results

3.1. Comparison methods

In this study, we compared the performance of AdFed in survival prediction with three commonly used classification methods (random forest (RF), XGBoost, and support vector machine (SVM)) and two FDL-based methods (FedProx and FedAvg) that integrate encrypted data. Considering that the survival prediction model constructed through direct integration of unencrypted cancer data exhibits superior accuracy compared to that built upon encrypted data integration, we conducted a comparison study to evaluate the advantages of integrating encrypted data using the following rules:

For RF, XGBoost, and SVM, we utilized datasets from the same cancer type collected across multiple institutions to construct separate prediction models and subsequently computed the average accuracy of the test dataset. For the FDL-based methods (FedProx, FedAvg, and AdFed), we incorporated encrypted data (excluding test data) to construct an integrated model for predicting the test data. When all datasets pertaining to the same type of cancer have been tested, the average accuracy serves as an indicator of the method's performance in predicting survival outcomes.

3.2. Prediction performance comparison

For a comprehensive comparison of different methods, we present the prediction accuracy using Accuracy, AUC, and F1 in Table 3 and Fig. 2. The three classification techniques (RF, XGBoost, and SVM) exhibit comparable predictive performance. As shown in Table 3, the traditional approaches (RF, SVM, XGBoost) achieved an average AUC value of 0.503 which is lower than that obtained by FDL-based methods (0.561). It is worth noting that the average AUC value of SVM is higher than that of FedProx, mainly due to the low AUC obtained by FedProx in HNSC (0.396). However, in other datasets, FedProx outperforms SVM in classification. These results suggest that without data integration, classification methods may have inferior prediction performance due to limited data availability compared to FDL-based methods with integrating encrypted data. Among the FDL-based methods, FedAvg achieves higher AUC values than FedProx but lower than AdFed. Notably, AdFed outperforms others with the highest performance score (ACC = 0.605, AUC = 0.588 and F1-score = 0.524).

3.3. Feature selection by AdFed

To assess the biological interpretability of the prediction model obtained by AdFed, we present the coefficients of the top ten survival-associated genes in four types of cancer in Fig. 3. Many of these identified genes have been validated to be linked with cancer progression. For instance, MMP12 has been identified as a negative prognostic marker in colon cancer [21]. Although there are no direct studies linking OXT to liver cancer, it is one of the key differential genes in non-alcoholic fatty liver disease [22]. S100A7 has been reported to regulate ovarian cancer cell metastasis and chemoresistance through the MAPK signaling pathway [23].

After obtaining the selected features by AdFed, we performed a weighted correlation network analysis (WGCNA) to eliminate irrelevant genes. As a case study, in Fig. 4A we conducted WGCNA on liver cancer data by using the R package *WGCNA* [24], resulting

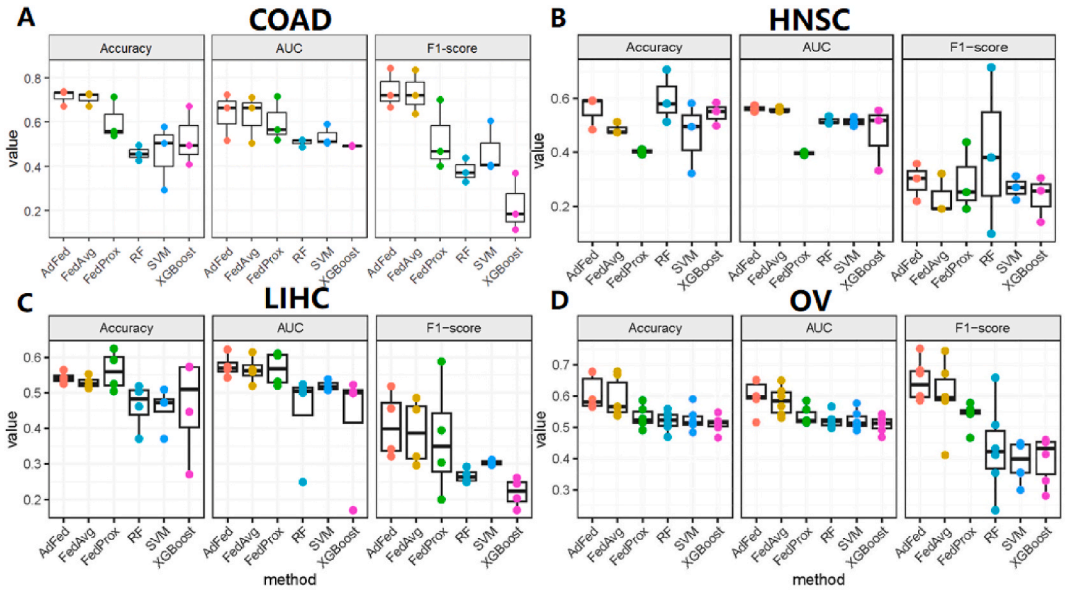


Fig. 2. The corresponding method performances for predicting cancer survival were obtained for different cancers: (A) colon cancer (COAD), (B) head and neck squamous cell carcinoma (HNSC), (C) liver cancer (LIHC), (D) ovarian cancer (OV).

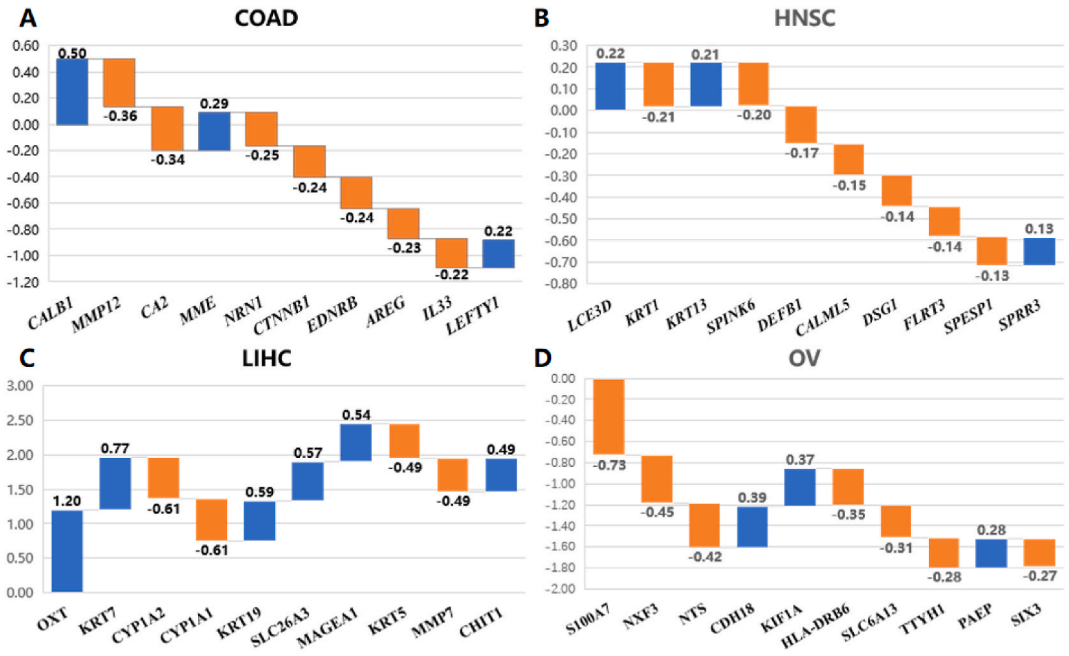


Fig. 3. The coefficients of the top 10 survival-related genes in four cancers obtained by AdFed: (A) colon cancer (COAD), (B) head and neck squamous cell carcinoma (HNSC), (C) liver cancer (LIHC), (D) ovarian cancer (OV).

in the identification of seven functional modules as shown in Fig. 4B. In WGCNA, we used an unsigned network and set the minimum number of genes in each module to 10. As a result, 51 genes with high GS scores (>0.7) were identified as candidate biomarkers for liver cancer. Among these 51 genes, the top 10 genes ranked based on the weights calculated by AdFed are shown in Fig. 4C.

To validate the interpretation of our experimental findings, we conducted a comprehensive investigation into their functional implications. Our results demonstrate that key genes have been confirmed to be associated with liver cancer, while the remaining ones have been reported to exhibit associations with various other types of cancers. Furthermore, the identified genes and elucidated pathways underscore the ability of AdFed to generate biologically meaningful predictions for cancer survival. In the literature review research, our findings revealed that half of the 10 genes (highlighted in bold in Table 4) have been validated to be linked with liver cancer, while the remaining ones are reported to be associated with other types of cancers. Furthermore, Table 3 presents the evidence regarding the functions of the selected genes. Using the 51 genes selected by AdFed and WGCAN, we conducted a KEGG pathway analysis to identify the enriched pathways with the *KOBAS* tool [25]. As depicted in Fig. 4D, the PI3K-Akt, Ras, and MAPK signaling pathways have been demonstrated to be associated with cancer prognosis [26]. Besides, these enriched pathways encompass crucial pathways that may impact cancer progression and drug response. In summary, the identified genes and pathways demonstrate that AdFed can produce biologically meaningful results in predicting cancer survival.

4. Discussion

In this study, we have introduced AdFed - an adaptive decentralized federal learning framework that leverages patient information from various organizations to construct cancer survival prediction models while addressing privacy protection concerns. Our experimental results demonstrate that AdFed utilizing distributed data outperforms the compared methods in cancer survival prediction. The results demonstrate that AdFed achieves a 2.7 % increase in AUC compared to the other FDL-based method. Additionally, AdFed can be combined with the regularization method to facilitate the identification of cancer-related genes by researchers. These results indicate that AdFed outperforms better than other federated-learning-based methods, and the interpretable algorithm can select biologically significant genes and pathways while ensuring the confidentiality and integrity of data.

Although the experimental results have demonstrated the accuracy and reliability of AdFed in cancer research, there remain certain issues that warrant further discussion. AdFed is a specific approach within the broader realm of federated learning. Compared with other federated learning-based approaches in the context of cancer research, AdFed employs robust privacy protection mechanisms, enabling collaborative model updates while preserving the confidentiality of individual data. It ensures privacy preservation even in the presence of malicious participants, safeguarding sensitive information throughout the computation process. AdFed adopts a decentralized approach by localizing data on individual devices or nodes, mitigating privacy risks associated with centralized storage or sharing. Moreover, AdFed facilitates training on a distributed network of numerous devices or nodes, effectively distributing computational load and ensuring scalability. However, if the client has a single point of failure, the central server becomes a potential single point of failure. In case the server becomes unavailable or compromised, it could disrupt the entire training process. AdFed

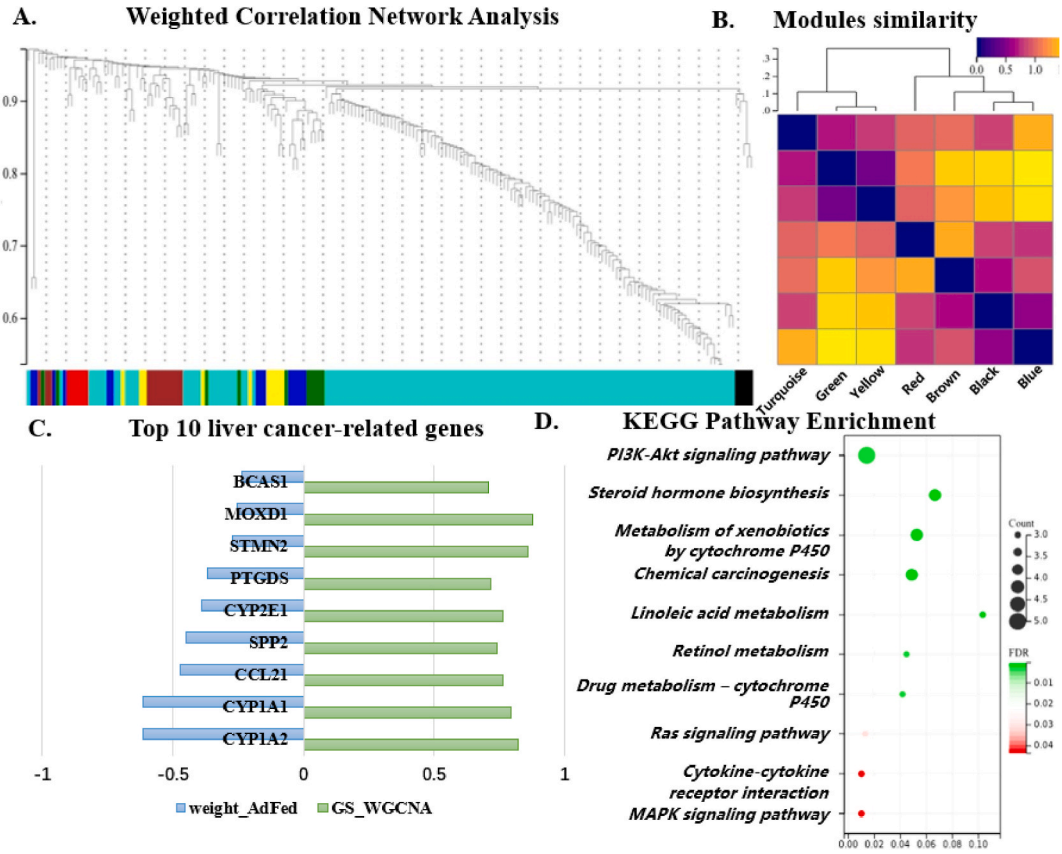


Fig. 4. (A) Hierarchical clustering result in weighted correlation network analysis (WGCNA) by using the liver cancer data. (B) Correlation matrix between different modules in WGCNA. (C) The top 10 genes selected by AdFed and WGCNA, ranked based on the importance calculated by AdFed. (D) The enriched liver cancer-related pathways obtained by KEGG pathway analysis.

Table 4
Identified biomarkers related to the survival of patients with liver cancer.

Rank	Gene	Importance_AdFed	GS_WGCNA	Function_Reference
1	**CYP1A2**	0.614	0.822	[27]
2	**CYP1A1**	0.612	0.795	[27]
3	CCL21	0.472	0.765	[28]
4	**SPP2**	0.448	0.742	[29]
5	**CYP2E1**	0.389	0.761	[30]
6	PTGDS	0.365	0.717	[31]
7	**STMN2**	0.274	0.859	[32]
8	MOXD1	0.253	0.877	[33]
9	BCAS1	0.236	0.707	[34]
10	COL10A1	0.227	0.830	[35]

necessitates sharing model updates with the central server. Although it does not involve sharing raw data, there might still be privacy concerns regarding sensitive medical information being shared. Additionally, AdFed involves dealing with devices or nodes that possess varying capabilities, network conditions, and data distributions; this can introduce challenges in achieving fair and representative training. Additionally, our methodology is not conducive to the secure and expeditious sharing of cancer imaging data, despite such images often containing valuable information. Thirdly, the genetic information provided by different institutions may be inconsistent. AdFed requires the same dimension of sample features. Therefore, we had to remove different gene features during data preprocessing. To address these limitations, in future work, we will further enhancing the transfer efficiency between models and updating the cancer survival prediction framework to enable the integration of patient imaging data in the future.

Consent for publication

All the authors listed have approved the manuscript.

Data availability

All the used data are collected from GEO (GSE-17538/38832/41613/42743/10141/14520/54236/14764/17260/18520/32062/63885, https://www.ncbi.nlm.nih.gov) and TCGA (TCGA-COAD/HNSC/LIHC/OV, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

Funding

This work was funded by the Jihua laboratory scienctific project (X210101UZ210) and the National Natural Science Foundation of China (62201150,62301006).

CRediT authorship contribution statement

Hua Chai: Writing – review & editing, Writing – original draft, Conceptualization. **Yiqian Huang:** Methodology, Data curation. **Lekai Xu:** Methodology, Formal analysis. **Xinpeng Song:** Methodology, Formal analysis. **Minfan He:** Formal analysis. **Qingyong Wang:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

References

[1] E.S. Wong, R.W. Choy, Y. Zhang, W.K. Chu, L.J. Chen, C.P. Pang, J.C. Yam, Global retinoblastoma survival and globe preservation: a systematic review and meta-analysis of associations with socioeconomic and health-care factors, Lancet Global Health (2022).
[2] J. Zhou, L. Li, L. Wang, X. Li, H. Xing, L. Cheng, Establishment of a SVM classifier to predict recurrence of ovarian cancer, Mol. Med. Rep. 18 (4) (2018) 3589–3598.
[3] M. Pieretti, C. Hopenhayn-Rich, N.H. Khattar, Y. Cao, B. Huang, T.C. Tucker, Heterogeneity of ovarian cancer: relationships among histological group, stage of disease, tumor markers, patient characteristics, and survival, Cancer Invest. 20 (1) (2002) 11–23.
[4] S. Wang, Z. Liu, Y. Rong, B. Zhou, Y. Bai, W. Wei, M. Wang, Y. Guo, J. Tian, Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer, Radiother. Oncol. 132 (2019) 171–177.
[5] H. Chai, X. Zhou, Z. Zhang, J. Rao, H. Zhao, Y. Yang, Integrating multi-omics data through deep learning for accurate cancer prognosis prediction, Comput. Biol. Med. 134 (2021) 104481.
[6] N.B. Truong, K. Sun, G.M. Lee, Y. Guo, Gdpr-compliant personal data management: a blockchain-based solution, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1746–1761.
[7] H. Li, F. Guo, W. Zhang, J. Wang, J. Xing, (a,k)-Anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems, J. Med. Syst. 42 (3) (2018) 1–9.
[8] H. Li, Y. Dai, X. Lin, Efficient e-health data release with consistency guarantee under differential privacy. 2015 17th International Conference on E-Health Networking, Application & Services (HealthCom): 2015, IEEE, 2015, pp. 602–608.
[9] J.L. Raisaro, G. Choi, S. Pradervand, R. Colsenet, N. Jacquemont, N. Rosat, V. Mooser, J.-P. Hubaux, Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy, IEEE/ACM transactions on computational biology and bioinformatics and bioinformatics 15 (5) (2018) 1413–1426.
[10] R.S. Antunes, C. André da Costa, A. Küderle, I.A. Yari, Eskofier B: federated learning for healthcare: systematic review and architecture proposal, ACM Transactions on Intelligent Systems and Technology 13 (4) (2022) 1–23.
[11] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, X. Jiang, Privacy-preserving patient similarity learning in a federated environment: development and analysis, JMIR medical informatics 6 (2) (2018) e7744.
[12] Y. Chen, X. Qin, J. Wang, C. Yu, W. Gao, Fedhealth: a federated transfer learning framework for wearable healthcare, IEEE Intell. Syst. 35 (4) (2020) 83–93.
[13] H. Elayan, M. Aloqaily, M. Guizani, 2021 International wireless communications and mobile computing (IWCMC): 2021. Deep Federated Learning for IoT-Based Decentralized Healthcare Systems, IEEE, 2021, pp. 105–109.
[14] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proceedings of Machine learning and systems 2 (2020) 429–450.
[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, y Arcas BA: communication-efficient learning of deep networks from decentralized data, In: Artificial intelligence and statistics: 2017: PMLR (2017) 1273–1282.
[16] Y. Li, R. Wang, Y. Li, M. Zhang, C. Long, Wind power forecasting considering data privacy protection: a federated deep reinforcement learning approach, Appl. Energy 329 (2023) 120291.
[17] Y. Li, X. Wei, Y. Li, Z. Dong, M. Shahidehpour, Detection of false data injection attacks in smart grid: a secure federated deep learning approach, IEEE Trans. Smart Grid 13 (6) (2022) 4862–4872.
[18] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.
[19] Y. Qu, M.P. Uddin, C. Gan, Y. Xiang, L. Gao, J. Yearwood, Blockchain-enabled federated learning: a survey, ACM Comput. Surv. 55 (4) (2022) 1–35.
[20] C. Xu, Y. Qu, Y. Xiang, L. Gao, Asynchronous federated learning on heterogeneous devices: a survey, Computer Science Review 50 (2023) 100595.
[21] F. Klupp, L. Neumann, C. Kahlert, J. Diers, N. Halama, C. Franz, T. Schmidt, M. Koch, J. Weitz, M. Schneider, Serum MMP7, MMP10 and MMP12 level as negative prognostic markers in colon cancer patients, BMC Cancer 16 (1) (2016) 1–9.
[22] M.R. Tavirani, M.R. Tavirani, M.Z. Azodi, ANXA2, PRKCE, and OXT are critical differentially genes in Nonalcoholic fatty liver disease, Gastroenterology and hepatology from bed to bench 12 (2) (2019) 131.
[23] M. Lin, B. Xia, L. Qin, H. Chen, G. Lou, S100A7 regulates ovarian cancer cell metastasis and chemoresistance through MAPK signaling and is targeted by miR-330-5p, DNA Cell Biol. 37 (5) (2018) 491–500.

[24] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (1) (2008) 1–13.
[25] C. Xie, X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C.Y. Li, L. Wei, Kobas 2.0: a web server for annotation and identification of enriched pathways and diseases, Nucleic Acids Res. 39 (Web Server issue) (2011) W316–W322.
[26] R. Gedaly, P. Angulo, J. Hundley, M.F. Daily, C. Chen, A. Koch, B.M. Evers, PI-103 and sorafenib inhibit hepatocellular carcinoma cell proliferation by blocking Ras/Raf/MAPK and PI3K/AKT/mTOR pathways, Anticancer Res. 30 (12) (2010) 4951–4958.
[27] J. Terashima, S. Goto, H. Hattori, S. Hoshi, M. Ushirokawa, K. Kudo, W. Habano, S. Ozawa, CYP1A1 and CYP1A2 expression levels are differentially regulated in three-dimensional spheroids of liver cancer cells compared to two-dimensional monolayer cultures, Drug Metabol. Pharmacokinet. 30 (6) (2015) 434–440.
[28] T.-L. Hwang, L.-Y. Lee, C.-C. Wang, Y. Liang, S.-F. Huang, C.-M. Wu, CCL7 and CCL21 overexpression in gastric cancer is associated with lymph node metastasis and poor prognosis, World J. Gastroenterol.: WJG 18 (11) (2012) 1249.
[29] Y. Tu, C. Chen, G. Fan, Association between the expression of secreted phosphoprotein-related genes and prognosis of human cancer, BMC Cancer 19 (1) (2019) 1–12.
[30] F. Webster, I.B. Lambert, C.L. Yauk, Adverse Outcome Pathway on Cyp2E1 Activation Leading to Liver Cancer, 2021.
[31] P. Jiang, Y. Cao, F. Gao, W. Sun, J. Liu, Z. Ma, M. Xie, S. Fu, SNX10 and PTGDS are associated with the progression and prognosis of cervical squamous cell carcinoma, BMC Cancer 21 (1) (2021) 1–14.
[32] H.-S. Lee, D.C. Lee, M.-H. Park, S.-J. Yang, J.J. Lee, D.M. Kim, Y. Jang, J.-H. Lee, J.Y. Choi, Y.K. Kang, STMN2 is a novel target of β-catenin/TCF-mediated transcription in human hepatoma cells, Biochem. Biophys. Res. Commun. 345 (3) (2006) 1059–1067.
[33] P. Shi, J. Xu, F. Xia, Y. Wang, J. Ren, P. Liang, H. Cui, MOXD1 knockdown suppresses the proliferation and tumor growth of glioblastoma cells via ER stress-inducing apoptosis, Cell Death Discovery 8 (1) (2022) 174.
[34] R. Correa, A. De Carvalho, N. Pinheiro, A. Simpson, S. De Souza, NABC1 (BCAS1): alternative splicing and downregulation in colorectal tumors, Genomics 65 (3) (2000) 299–302.
[35] H. Huang, T. Li, G. Ye, L. Zhao, Z. Zhang, D. Mo, Y. Wang, C. Zhang, H. Deng, G. Li, High expression of COL10A1 is associated with poor prognosis in colorectal cancer, OncoTargets Ther. (2018) 1571–1581.